一种高效的 CD-CAT 在线标定新方法：基于熵的信息增益与 EM 视角*

谭青蓉[1] 汪大勋[1] 罗芬[1] 蔡艳[1] 涂冬波[1]

([1]江西师范大学心理学院，南昌 330022)

摘 要 项目增补(Item Replenishing)对认知诊断计算机自适应测验(CD-CAT)题库的维护有着至关重要的作用，而在线标定是一种重要的项目增补方式。基于数据挖掘中特征选择(Feature Selection)的思路，提出一种高效的基于熵的信息增益的在线标定方法(记为IGEOCM)，该方法利用被试在新旧题上的作答联合估计新题的Q矩阵和项目参数。研究采用Monte Carlo模拟实验验证所开发新方法的效果，并同时与已有的在线标定方法SIE (Chen et al., 2015)、SIE-R-BIC和RMSEA-N (谭青蓉，2019)进行比较。结果表明：新开发的IGEOCM在各实验条件下均具有较好的项目标定精度和项目估计效率，且整体上优于已有的SIE等方法；同时，IGEOCM标定新题所需的时间低于SIE等方法。总之，研究为CD-CAT题库中项目的增补提供了一种更为高效、准确的方法。

关键词 认知诊断计算机自适应测验，项目增补，在线标定，Q矩阵，熵的信息增益

分类号 B841

1 引言

测评技术与计算机技术的持续发展，使得大众不仅追求测验的效率，更追求综合性的测验结果，而不仅仅是笼统的测验总分。人们渴望获取详实且全面的测验结果，使其能根据该结果对自身在所测内容领域上的强弱进行系统评估，了解其需改进或完善的地方，从而制定进一步的学习计划。认知诊断计算机自适应测验(Cognitive Diagnostic Computerized Adaptive Testing, CD-CAT)是认知诊断

* 收稿日期：2020-11-30
国家自然科学基金(31760288, 31660278, 31960186)
通信作者：涂冬波，E-mail: tudongbo@aliyun.com

(Cognitive Diagnostic, CD)与计算机自适应测验(Computerized Adaptive Testing, CAT)相结合的产物，其在提高测验效率和准确性的同时，可为被试提供在所测内容领域上优缺点的详细诊断(Wang, 2013; Weiss, 1982)。因此，可根据被试的诊断结果对其薄弱知识点进行针对性地教学补救，较好地满足了当今大众对于高效且周密的测验的需求，有着广泛的应用前景(Leighton et al., 2004; Liu et al., 2013)。

CD-CAT 使用的前提是已建构好的题库。但是题库中的部分题目会随时间的流逝过度曝光或变得过时，这时需使用新题对这些题目进行替换或者增补(Chen, 2017)。具体来说，需邀请有经验的领域专家和心理测量学家根据诊断目的编制新题，然后估计新题的参数，并将其与题库中的旧题置于同一量尺之上。在线标定技术是传统 CAT 中一种有效的项目增补方法，它是指在测验过程中，让被试同时作答新题与旧题，然后根据其作答标定新题参数的过程，且施测者需告知被试他们作答的部分项目将不用于最终能力的评估(陈平, 辛涛, 2011a)。相比于传统的项目增补方法，在线标定技术的优点在于：(1)无需复杂的事后等值技术便可将新旧题的参数置于同一量尺之上(Chen & Wang, 2015)；(2)无需外部标定研究便能在估计被试能力的同时标定新题的参数，可节省大量人力和物力；(3)相同的测量模式使得被试在作答新旧题时具有相同的动机(Chen et al., 2012)。迄今为止，在单维计算机自适应测验(Unidimensional Computerized Adaptive Testing, UCAT)和多维计算机自适应测验(Multidimensional Computerized Adaptive Testing, MCAT)领域，研究者已推荐了多种高效的在线标定方法(Chen, 2017)。在 UCAT 中，Stocking (1988)提出方法 A (Method A)和方法 B (Method B)，Wainer 和 Mislevy (1990)推荐一个 EM 循环的边际极大似然估计方法(OEM)，随后 Ban 等人(2001)提出多个 EM 循环的边际极大似然估计方法(MEM)以及 BILOG/先验方法(BILOG/Prior method)。此外，为克服 Method A 方法将估计能力值当做被试能力真值的理论缺陷，陈平(2016)提出了 FFMLE-Method A 和 ECSE-Method A 方法。在 MCAT 中，Chen 等人(2017)对 Method A，OEM 和 MEM 方法进行拓展，称其为 M-Method A，M-OEM 和 M-MEM。且在 M-OEM 和 M-MEM 方法的基础上推荐 M-OEM-BME 和 M-MEM-BME 方法以用于 MCAT 中项目参数的标定(Chen, 2017)。

然而，目前CD-CAT中关于在线标定方法的研究较少，主要包含了两大类。

第一类方法主要有Chen等人(2012)提出的CD-Method A, CD-OEM与CD-MEM方法，其基于Method A, OEM与MEM提出。这类方法在标定新题时，假设新题的*Q*矩阵已知，仅标定新题的项目参数。事实上，*Q*矩阵作为认知诊断的核心成分，在多数情况下都是未知的。实际中，*Q*矩阵多由内容领域专家和测量学专家共同界定，需耗费大量的人力和物力，且由专家界定的*Q*矩阵容易受到主观因素的影响而造成界定错误。而*Q*矩阵的错误界定最终会影响项目参数的估计精度与被试的分类正确性(de la Torre & Chiu, 2016; Rupp & Templin, 2008)。因此，第二类在线标定方法应运而生，其同时标定新题的*Q*矩阵与项目参数，以期减少项目标定所耗费的人力物力，提高项目标定效率。陈平和辛涛(2011b)提出的联合估计算法(Joint Estimation Algorithm, JEA)，Chen等人(2015)提出的SIE (Single-Item Estimation)方法以及谭青蓉(2019)提出的SIE-R-BIC和RMSEA-N等方法均属于该类方法。JEA方法借鉴项目反应理论(Item Response Theory, IRT)中被试参数与项目参数的联合极大似然估计(Joint Maximum Likelihood Estimation, JMLE)思路，将CD-CAT中被试的属性掌握模式估计值视为被试属性掌握模式真值，然后基于被试属性掌握模式估计值以及被试在新题上的作答使用极大似然估计(Maximum Likelihood Estimation, MLE)的方法来联合估计新题的*Q*矩阵和项目参数。不同于JEA方法，SIE使用属性掌握模式的后验分布来代替属性掌握模式估计值，计算每一个被试的后验预测分布，然后使用MLE来估计新题的*Q*矩阵。与此同时，SIE方法中运用EM算法来估计新题的项目参数。SIE-R-BIC方法是在SIE方法的基础上提出，其标定新题时充分利用了题库中已有项目的信息，而RMSEA-N方法通过评估观察作答分布与期望作答分布间的一致性来标定新题(谭青蓉, 2019)。相比于JEA方法，SIE、SIE-R-BIC和RMSEA-N方法在*Q*矩阵标定精度上有一定的提升，而在标定效率上，各方法均耗时较长，新题标定效率相对较低。因此，在CD-CAT情境下，开发能提升新题标定精度和标定效率的方法是极为必要的。

数据挖掘作为数据库和人工智能领域研究的热点问题，其面临的首要问题是如何才能从海量数据中获得有效信息，从而达到数据信息的高效利用(Chandrashekar & Sahin, 2014)。特征选择(Feature Selection)是有效的解决方法之一，其可通过删除数据中冗余或无关的特征，从海量的数据中选择最为有效的特征集，以达到提高分类准确率以及效率的目的(Guyon & Elisseeff, 2003)。特征选

择过程中极为重要的一环是特征选择标准，其通过衡量特征与分类之间的关系来删除数据中的无关特征。特征选择中使用信息增益、互信息、归一化互信息以及条件互信息等作为特征选择标准，这类标准通过评估特征的分类准确性来选择最佳的特征(Fleuret, 2004; Lee et al., 2012; Hoque et al., 2014; Pereira et al., 2015)。特征对被试的分类越精确，则选择该特征的可能性越高，若特征对被试的分类相当于随机水平，则选择该特征的可能性越低。

受数据挖掘中特征选择的启发，提出如下逻辑假设：在CD-CAT中标定新题时，可利用特征选择方法来标定新题的Q矩阵，并基于该Q矩阵来估计新题项目参数。将新题所有可能的q向量视为待选择的特征，在被试属性掌握模式已知的情况下，通过特征选择标准评估每一个可能q向量对被试分类的效果，然后选择能使特征选择标准最佳的q向量作为新题的q向量。基于该假设，研究提出一种新的CD-CAT在线标定方法，该方法基于特征选择方法联合在线标定新题的Q矩阵和项目参数(该方法的基本过程、思路及公式等将在文章第3部分详细介绍)，以期为CD-CAT在线标定提供新的视角及新的方法，从而进一步推动认知诊断尤其是CD-CAT在实践中的发展与应用。

2 已有在线标定方法

目前，CD-CAT 中同时标定新题 Q 矩阵和项目参数的在线标定方法主要有JEA (陈平，辛涛, 2011b)，SIE (Chen et al., 2015)，SIE-R-BIC 和 RMSEA-N 方法(谭青蓉, 2019)。SIE 方法基于 JEA 方法在决定型输入噪音与门模型(the Deterministic Input, Noisy and Gate Model, DINA; Junker & Sijtsma, 2001)下提出，其标定新题时考虑了被试属性掌握模式的估计误差，在标定新题 Q 矩阵和项目参数时充分利用被试的属性掌握模式后验分布。

SIE 方法标定新题时包含了 Q 矩阵标定和项目参数标定的两个部分。对于新题 Q 矩阵的标定，首先基于被试在旧题上的作答计算作答了新题 j 的被试的属性掌握模式后验分布。随后，根据被试属性掌握模式后验分布及每种属性掌握模式在 q 向量为q_j的新题 j 上的正确作答概率计算具有某一特定作答 R_{ij} 的被试 i 的后验预测分布：

$$P_i(q_j, g_j, s_j) = P(R_{ij}=1|q_j, g_j, s_j) = \sum_{c=1}^{2^K} \pi_i(\alpha_c) P_j(q_j, g_j, s_j, \alpha_c), \quad (1)$$

其中 K 为测验测量的属性个数，$\pi_i(\alpha_c)$表示被试 i 的属性掌握模式为α_c的概率，其基于被试 i 在旧题上的作答计算，$P_j(q_j, g_j, s_j, \alpha_c)$表示 DINA 模型下属性掌握模式为$\alpha_c$的被试在项目 j 上的正确作答概率。最后，结合被试后验预测分布及其在新题 j 上的作答 R_{ij} 构建似然并最大化似然函数来估计新题的 q 向量，其表达式如下：

$$\hat{q}_j = \underset{q_j^* \in Q_j}{\operatorname{argmax}} L_j(q_j^*, g_j, s_j) = \operatorname{argmax}\left\{\prod_{i=1}^{n_j} P_i(q_j^*, g_j, s_j)^{R_{ij}} [1 - P_i(q_j^*, g_j, s_j)]^{1-R_{ij}}\right\}, (2)$$

其中$Q_j = 2^K - 1$表示新题 j 所有可能 q 向量的集合。此外，SIE 方法使用 EM 算法来估计新题的项目参数。

SIE-R-BIC 方法在 SIE 方法的基础上考虑了模型的复杂性，其估计新题 Q 矩阵时构建了 BIC 指标并通过最小化 BIC 指标来估计新题 q 向量，表达式如下所示：

$$\hat{q}_j = \underset{q_j^* \in Q_j}{\operatorname{argmin}} BIC_j(q_j^*, g_j, s_j) = \underset{q_j^* \in Q_j}{\operatorname{argmin}} [L_j(q_j^*, g_j, s_j) + \lambda \log(n_j)], \quad (3)$$

其中$\lambda \log(n_j)$表示模型复杂性的惩罚，λ 表示自由参数的个数，n_j表示作答新题 j 的被试人数。与此同时，SIE-R-BIC 方法在标定新题项目参数时利用了题库中已有项目的信息，也即将题库中和新题具有相同 q 向量的旧题的项目参数均值作为新题的项目参数初始值。RMSEA-N 方法中项目参数的标定与 SIE-R-BIC 方法一致，但其通过评估观察作答分布与期望作答分布间的一致性来标定新题的 Q 矩阵。具体来说，选择能使观察作答分布与期望作答分布间一致性程度最高的 q 向量作为新题 j 的估计 q 向量，其公式如下：

$$\hat{q}_j = \underset{q_j^* \in Q_j}{\operatorname{argmin}} RMSEA - N_j(q_j^*, g_j, s_j) = \underset{q_j^* \in Q_j}{\operatorname{argmin}} \sqrt{\sum_{c=1}^{2^K} p(\alpha_c) [P_N_{expected}(\alpha_c) - P_N_{observed}(\alpha_c)]^2}, (4)$$

其中 $p(\alpha_c)$ 表示第 c 个属性掌握模式α_c的被试边际概率，$P_N_{expected}(\alpha_c)$ 和 $P_N_{observed}(\alpha_c)$分别表示第 c 个属性掌握模式α_c下标准化的期望正确作答概率和观

察正确作答概率。

3 基于熵的信息增益在线标定方法(IGEOCM)

3.1 特征选择方法及基于熵的信息增益

数据挖掘中，特征选择的目的之一在于选择对数据具有高区分能力的特征，若基于某一特征的分类与随机分类的结果大同小异，则说明这一特征对于数据的分类效果较小(李航，2012)。基于熵的信息增益(Information Gain of Entropy-based, IGE)是特征选择中的一个特征选择准则，某一特征所具有的基于熵的信息增益值越大，则其对于数据的分类能力越强(Pereira et al., 2015)。基于熵的信息增益选择最优特征的过程如下：

(1)首先，确定数据集 R 以及对该数据集进行分类的特征。

(2)然后，计算数据集 R 的熵

$$E(R) = -\sum_{x=0}^{1} \frac{n_x}{n} \log \frac{n_x}{n}, \tag{5}$$

上式中，n 为数据集 R 的样本量，x 表示数据集 R 中的类别，n_x 为数据集 R 中属于第 x 个类别的样本量。熵用于评估数据集 R 的不确定性程度，其值越大，数据集 R 的不确定性程度越大。不确定性程度指数据集 R 中被试的一致性程度，若数据集 R 中的被试均属于同一个类别，则不确定性程度最低。

(3)随后，计算某一特征 A 对数据集 R 的条件熵

$$E(R|A) = -\sum_{h=1}^{H} \frac{n_h}{n} \left(\sum_{x=0}^{1} \frac{n_{hx}}{n_h} \log \frac{n_{hx}}{n_h} \right), \tag{6}$$

其中，H 为特征 A 的取值个数，n_h 表示数据集 R 中属于第 h 个类别的被试数量，n_{hx} 表示数据集 R 中第 h 个子类别下，被试属于第 x 个类别的数量。条件熵用于评估给定某一特征(A)的情况下，数据集 R 的不确定性程度。与熵一致，条件熵的值越大，数据集 R 的不确定性程度越大，则基于特征 A 的分类效果越差。

(4)最后，计算熵的信息增益值

$$g(R, A) = E(R) - E(R|A). \tag{7}$$

熵的信息增益$g(R,A)$为熵与条件熵之差，其表示在给定特征 A 的信息的情况下，数据集 R 的不确定性减少的程度。其值越大，说明基于特征 A 的分类效果越好。

(5)对于所有特征，重复(3)和(4)，比较各特征的熵信息增益值。选择具有最大的熵信息增益值的特征作为最优特征。

熵的信息增益的大小取决于数据集 R 的熵($E(R)$)和特征 A 对数据集 R 的条件熵($E(R|A)$)。由公式(5)可看出，数据集 R 的熵值的计算与特征无关，换句话说，所有特征下数据集 R 的熵($E(R)$)均保持不变。因此，基于熵的信息增益选择特征本质上是基于条件熵选择特征，某一特征对数据集 R 的条件熵越小，则该特征的分类效果越好，该特征更有可能是数据集的最优特征。

新题 j 的 q 向量估计可视为一个特征选择问题，即从所有可能的 q 向量中为新题 j 选择一个最佳 q 向量。将被试在新题 j 上的作答看作数据集 R，新题 j 所有可能的 q 向量看作特征，基于 q 向量和被试的估计属性掌握模式对被试进行分类，则能使被试分类的不确定性程度达到最低的 q 向量为作答数据集 R 的最优特征，因此可选择该 q 向量作为新题 j 的估计 q 向量。基于该思路，提出新的在线标定方法—基于熵的信息增益的在线标定方法(Information Gain of Entropy-based Online Calibration Method，IGEOCM)，该方法使用熵的信息增益来标定新题的 Q 矩阵，同时使用 EM 算法来标定新题的项目参数。

3.2 基于熵的信息增益在线标定方法开发

DINA 作为广泛应用的认知诊断模型之一，在每个项目上均只有失误参数和猜测参数这两个简单且易于解释的项目参数，且常被用于 CD-CAT 题库的构建及在线标定(Junker & Sijtsma, 2001; Liu et al., 2013)。为了便于说明问题以及与国外同类方法(SIE 方法)进行比较，以 DINA 模型为例来说明基于熵的信息增益的在线标定方法(IGEOCM)标定新题的基本思路及其过程。

3.2.1 IGEOCM 中的 Q 矩阵标定

当新题所测量的属性个数 K 已知时，新题 j 所有可能的 q 向量个数为2^K-1，其中不包含元素全为 0 的向量。从特征选择的视角，新题 j 的 q 向量估计便是从2^K-1种可能 q 向量中选择最合适的一个 q 向量作为新题 j 的估计 q 向量。IGEOCM 中基于熵的信息增益这一特征选择准则来估计新题 j 的 q 向量，其表达式如下所示：

$$\begin{aligned} g(R_j,q_j) &= E(R_j) - E(R_j|q_j) \\ &= \left(-\sum_{x=0}^{1}\frac{n_{jx}}{n_j}\log\frac{n_{jx}}{n_j}\right) - \left(-\sum_{h=0}^{1}\frac{n_{jh}}{n_j}\left(\sum_{x=0}^{1}\frac{n_{jhx}}{n_{jh}}\log\frac{n_{jhx}}{n_{jh}}\right)\right), \end{aligned} \quad (8)$$

其中，$R_j=(R_{1j},\cdots,R_{n_jj})$表示$n_j$个被试在新题$j$上的作答反应向量，即新题$j$上的被试作答数据集合。$n_j$表示作答了新题$j$的被试人数，$x$表示被试在项目$j$上的得分，二级计分下，$x$=0或$x$=1。$n_{jx}$表示作答了新题$j$的$n_j$个被试中在新题$j$上得分为$x$的被试人数。$h$表示基于$q$向量对被试进行分类的类别。DINA模型中，基于被试的属性掌握模式与项目的q向量可将被试划分为2个类别(h=1或h=0)，即掌握组与非掌握组。掌握组被试掌握了项目测量的所有属性，非掌握组被试在项目所测量的属性中至少有一个属性未掌握。n_{jh}表示作答新题j的n_j个被试中属于第h个类别的被试人数。n_{jhx}表示作答新题j且属于第h个类别的n_{jh}个被试中在新题j上得分为x的被试人数。对于新题j的所有可能q向量，作答新题j的被试人数(n_j)以及每个被试在新题j上的得分x都是不变的。因此，基于熵的信息增益标定新题q向量的本质在于，在新题j的q向量未知的情况下，选择能使条件熵$E(R_j|q_j)$最小的q向量作为新题j的估计q向量，其表达式为：

$$\hat{q}_j = \underset{q_j^* \in Q_j}{\operatorname{argmin}}\left(E(R_j|q_j^*)\right). \quad (9)$$

在被试属性掌握模式已知的情况下(即CD-CAT中基于被试在旧题上的作答估计属性掌握模式)，若新题j的q向量正确且被试在新题j上的作答不存在失误和猜测，那么基于正确q向量分类后掌握组中的所有被试在新题j上的观察得分都应为1，而非掌握组中的被试在新题j上的观察得分都应为0。此时，掌握组与非掌握组中的被试都具有高度一致性，不确定程度最小，所获得的条件熵$E(R_j|q_j)$最小，

信息增益$g(R_j,q_j)$最大，因此正确q向量的分类效果最好。若被试的属性掌握模式为均匀分布且被试在新题j上的作答不存在失误和猜测，新题j的q向量正确和错误情况下$E(R_j|q_j)$和$g(R_j,q_j)$的变化如表1所示。

表1 不同q向量下$E(R_j|q_j)$和$g(R_j,q_j)$的计算

q 向量		掌握组	非掌握组	$E(R_j\|q_j)$	$g(R_j,q_j)$
$q_j^{correct}=[100]$	属性掌握模式	[100] [110] [101] [111]	[000] [101] [001] [011]	0	0.690
	被试数目	$n_j/2$	$n_j/2$		
	正确作答比	1	0		
	错误作答比	0	1		
$q_j^{incorrect}=[011]$	属性掌握模式	[011] [111]	[000] [100] [010] [001] [110] [101]	0.690	0.003
	被试数目	$n_j/4$	$3n_j/4$		
	正确作答比	0.500	0.500		
	错误作答比	0.500	0.500		

IGEOCM 中通过最小化条件熵（$\hat{q}_j=\underset{q_j^*\in Q_j}{\operatorname{argmin}}\left(E(R_j|q_j^*)\right)$）估计新题$j$的$q$向量的合理性证明如下：

$$E(R_j|q_j)=\left(-\sum_{h=0}^{1}\frac{n_{jh}}{n_j}\left(\sum_{x=0}^{1}\frac{n_{jhx}}{n_{jh}}\log\frac{n_{jhx}}{n_{jh}}\right)\right)$$
$$=-\left[\frac{n_{j0}}{n_j}\left(\frac{n_{j00}}{n_{j0}}\log\frac{n_{j00}}{n_{j0}}+\frac{n_{j01}}{n_{j0}}\log\frac{n_{j01}}{n_{j0}}\right)+\frac{n_{j1}}{n_j}\left(\frac{n_{j10}}{n_{j1}}\log\frac{n_{j10}}{n_{j1}}+\frac{n_{j11}}{n_{j1}}\log\frac{n_{j11}}{n_{j1}}\right)\right], \quad (10)$$

在 DINA 模型下，$\hat{s}_j=\frac{n_{j10}}{n_{j1}}$（$n_{j1}$表示作答新题$j$的$n_j$个被试中属于掌握组的被试人数，$n_{j10}$表示作答新题$j$且属于掌握组的$n_{j1}$个被试中在新题$j$上得分为 0 的被试人数），$\hat{g}_j=\frac{n_{j01}}{n_{j0}}$（$n_{j0}$表示作答新题$j$的$n_j$个被试中属于非掌握组的被试人数，$n_{j01}$表示作答新题$j$且属于非掌握组的$n_{j0}$个被试中在新题$j$上得分为 1 的被试人数），将其代入$E(R_j|q_j)$，可得

$$E(R_j|q_j) = -\left[\frac{n_{j0}}{n_j}\left(\left(1-\hat{g}_j\right)\log\left(1-\hat{g}_j\right)+\left(\hat{g}_j\right)\log\left(\hat{g}_j\right)\right)+\frac{n_{j1}}{n_j}\left(\left(\hat{s}_j\right)\log\left(\hat{s}_j\right)+\left(1-\hat{s}_j\right)\log\left(1-\hat{s}_j\right)\right)\right], \quad (11)$$

对$E(R_j|q_j)$函数求关于$\hat{s}_j$和$\hat{g}_j$的偏导数，并令其等于 0，

$$\frac{\partial E(R_j|q_j)}{\partial \hat{s}_j} = -\frac{n_{j1}}{n_j}\left(\log\frac{\hat{s}_j}{1-\hat{s}_j}\right)=0,\quad \frac{\partial E(R_j|q_j)}{\partial \hat{g}_j} = -\frac{n_{j0}}{n_j}\left(\log\frac{\hat{g}_j}{1-\hat{g}_j}\right)=0. \quad (12)$$

通过代数运算可得$\hat{g}_j=\hat{s}_j=\frac{1}{2}$。由于

$$\frac{\partial^2 E(R_j|q_j)}{\partial \hat{s}_j^2} = -\frac{n_{j1}}{n_j}\left(\frac{1}{\hat{s}_j}\times\frac{1}{1-\hat{s}_j}\right)<0,\quad \frac{\partial^2 E(R_j|q_j)}{\partial \hat{g}_j^2} = -\frac{n_{j0}}{n_j}\left(\frac{1}{\hat{g}_j}\times\frac{1}{1-\hat{g}_j}\right)<0, \quad (13)$$

则$E(R_j|q_j)$在$\hat{g}_j=\hat{s}_j=\frac{1}{2}$处取得最大值。当$\hat{s}_j,\hat{g}_j\in(0,0.5)$时，$E(R_j|q_j)$单调递增。

另外，根据 Yu 和 Cheng (2020)的研究，当$n_j\to\infty$，属性掌握模式已知，且$s_j,g_j\in(0,0.5)$时，以下等式成立：

$$\hat{s}_j(q_j^{incorrect}) \geqslant \hat{s}_j(q_j^{correct}),\ \hat{g}_j(q_j^{incorrect}) \geqslant \hat{g}_j(q_j^{correct}). \quad (14)$$

因此，在被试的属性掌握模式以及被试在项目j上的作答已知，$\hat{s}_j,\hat{g}_j\in(0,0.5)$的情况下，新题$j$的$q$向量可通过最小化条件熵来估计：$\hat{q}_j=\underset{q_j^*\in Q_j}{\mathrm{argmin}}\left(E(R_j|q_j^*)\right)$。与此同时，附录中【例 1】通过举例进一步说明了 IGEOCM 最小化条件熵估计新题j的q向量的合理性。

3.2.2 IGEOCM 中的项目参数标定

IGEOCM方法中使用EM算法来估计新题的项目参数，EM算法在每一次迭代中都包含期望步骤(Expectation Step, E-step)和最大化步骤(Maximization Step, M-step)两步(Chen et al., 2015)。在E-step中，首先基于被试i在新题j上的作答R_{ij}计算每个被试的后验分布，其公式如下：

$$Post_{ij}(\alpha_c) = \frac{\pi_i(\alpha_c)P_j(q_j,g_j,s_j,\alpha_c)^{R_{ij}}[1-P_j(q_j,g_j,s_j,\alpha_c)]^{1-R_{ij}}}{\sum_{c=1}^{2^K}\pi_i(\alpha_c)P_j(q_j,g_j,s_j,\alpha_c)^{R_{ij}}[1-P_j(q_j,g_j,s_j,\alpha_c)]^{1-R_{ij}}}. \quad (15)$$

然后，基于n_j个被试在新题j上的作答向量R_j和每个被试属性掌握模式的后验分布，假设n_j个被试在新题j上的作答彼此独立，可构建对数边际似然函数如下：

$$L(q_j,g_j,s_j)=\prod_{i=1}^{n_j}\sum_{c=1}^{2^K}Post_{ij}(\alpha_c)\left[(R_{ij}\log P_j(q_j,g_j,s_j,\alpha_c))+(1-R_{ij})\log(1-P_j(q_j,g_j,s_j,\alpha_c))\right]. \quad (16)$$

M-step 的目的在于最大化公式(16)以估计新题 j 的失误参数 s_j 和猜测参数 g_j。EM 算法依次迭代 E-step 和 M-step 直到满足预先设定的收敛标准。

上述两个部分为 IGEOCM 对新题 Q 矩阵和项目参数的标定，其标定新题的具体步骤如下：

步骤 1：新题 q 向量估计。对于新题 j，基于作答了新题 j 的被试的属性掌握模式估计值及其在新题 j 上的作答数据，计算每一个可能 q 向量下作答数据集 R_j 的条件熵 $E(R_j|q_j)$，选择最小 $E(R_j|q_j)$ 值对应的 q 向量作为新题 j 的估计 q 向量。

步骤 2：新题项目参数估计。将步骤 1 中的估计 q 向量作为新题 j 的真实 q 向量，基于作答了新题 j 的被试的属性掌握模式后验分布及其在新题 j 上的作答，使用 EM 算法估计新题的失误参数和猜测参数。新题 j 标定完成。

步骤 3：对于所有待标定的其他新题，重复步骤 1 和步骤 2 可获得新题的 Q 矩阵估计值和项目参数(失误参数和猜测参数)估计值，直到所有新题标定完成。

IGEOCM 是基于特征选择的视角提出的在线标定新方法。该方法的优点在于仅需获得被试的属性掌握模式估计值以及被试在新题上的作答便能估计新题的 Q 矩阵，是一种非参数化的方法，简单易懂且无需复杂的计算。此外，IGEOCM 将基于非参数化方法估计的 q 向量作为新题的真实 q 向量直接标定新题的项目参数，不论新题可能 q 向量的多少，IGEOCM 均只需估计一个已确定 q 向量下的项目参数，可有效节约项目标定的时间，改善新题标定的效率。这不同于 SIE 方法，其需估计所有可能 q 向量下的项目参数，标定新题的时间长，标定新题的效率低。

4 研究 1：IGEOCM 和已有在线标定方法性能及其精度验证

4.1 实验设计

研究 1 旨在考查 IGEOCM 在不同标定样本(40、80、120、160、200)、属性掌握模式分布(均匀分布、高阶分布、多元正态分布)和被试作答新题个数 D (4、

6、8)下标定新题的效果，并将其与 SIE、SIE-R-BIC 和 RMSEA-N 方法进行比较。标定样本指作答了新题 j 的被试人数 $n_j=(N\times D)/m$，其中，N 为参与 CD-CAT 的被试总人数，D 为每个被试作答新题的个数，m 为待标定的新题个数(Chen et al., 2015)。选择 SIE，SIE-R-BIC 和 RMSEA-N 方法作为比较方法，主要原因在于其新题标定精度略优于 JEA 方法，具有一定的代表性。研究 1 为四因素实验设计，共 $5\times3\times3\times4=180$ 种模拟实验条件，每种实验条件重复实验 500 次以减少随机误差。

4.1.1 被试与题库模拟

标定样本共 5 个水平，$n_j=40, 80, 120, 160$ 和 200，被试属性掌握模式分别从均匀分布、高阶分布和多元正态分布 $MVN(0,\Sigma)$ 中产生。在均匀分布中，被试的属性掌握模式从所有可能的属性掌握模式中以均匀的概率产生；在高阶分布中，被试 i 是否掌握第 k 个属性与被试 i 的一般潜在能力 θ_i 有关，能力为 θ_i 的被试 i 掌握第 k 个属性的概率为

$$P(\alpha_{ik}|\theta_i)=\frac{\exp\left(\lambda_{1k}(\theta_i-\lambda_{0k})\right)}{1+\exp\left(\lambda_{1k}(\theta_i-\lambda_{0k})\right)}, \tag{17}$$

其中，λ_{0k} 和 λ_{1k} 为结构参数，$\lambda_{1k}>0$。研究中设置 $K=6$，$\lambda_0=(-1, -0.6, -0.2, 0.2, 0.6, 1)$，且对所有属性 k 均有 $\lambda_{1k}=1.5$，被试 i 的能力值从 $N(0,1)$ 中产生(de la Torre & Chiu, 2016)；在多元正态分布中，属性间的相关设置为 0.5 (J. Chen, 2017)。

题库模拟包含项目参数(失误参数 s 和猜测参数g)的模拟和项目 Q 矩阵的模拟。题库中共包含300个题目，每个题目最多测量3个属性，且题库中测量1、2和3个属性的项目均设置为100题。测验测量属性的总个数 $K=6$，则共有63种可能的项目 q 向量，其中测量1个属性的项目 q 向量个数为6，测量2个属性的项目 q 向量个数为15，测量三个属性的项目 q 向量个数为20。将测量1个属性的6个项目 q 向量重复16次并从其中额外抽取4个项目 q 向量，测量2个属性的15个项目 q 向量重复6次并

从其中额外抽取10个项目q向量，测量3个属性的20个项目q向量重复5次，构成300×6的临时测验Q矩阵。最后，对临时Q矩阵中的所有行随机排序以获得最后的Q矩阵。每一个项目的失误参数s和猜测参数g均从$U\,(0.05,0.25)$随机抽取。

4.1.2 新题模拟

新题的模拟包括新题失误参数s和猜测参数g的模拟以及新题Q矩阵的模拟。研究中，令需标定的新题个数$m=24$，因此新题的Q矩阵是一个24×6的矩阵。新题测验Q矩阵及其失误参数s和猜测参数g的模拟均与题库的模拟保持一致。

4.1.3 CD-CAT 模拟与新题标定

研究使用定长的终止规则，每个被试均作答 20 个旧题和 D 个新题($D=4,6,8$三个水平)。CD-CAT 的模拟过程如下：

测验开始时，由于对被试的情况一无所知，因此(1)随机从题库中抽取一个项目作为被试的初始作答项目；(2)模拟当前被试在项目上的作答，并通过被试在已作答项目上的作答使用 MLE 估计被试的属性掌握模式；(3)使用后验加权 KL (Posterior-Weighted Kullback-Leibler, PWKL; Cheng, 2009)选题策略从剩余题库中挑选最适合被试当前属性掌握模式估计值的项目作为被试的下一个作答项目。重复步骤(2)和(3)直到测验长度达到预先指定的标准。

在 CD-CAT 模拟过程中，随机从待标定的 24 个新题中抽取 D 个新题并将其置于被试测验过程的随机位置。CD-CAT 测验结束后，基于被试的属性掌握模式估计值，属性掌握模式后验分布及被试在新题上的作答，分别使用 IGEOCM、SIE、SIE-R-BIC 和 RMSEA-N 方法标定新题的 Q 矩阵和项目参数。

4.1.4 评价标准

属性向量正确估计率(Attribute Vector Correct Estimation Rate, AVCER) AVCER用于评估新题Q矩阵的估计正确率，其表达式为：

$$AVCER=\frac{1}{500\times m}\sum_{r=1}^{500}\sum_{j=1}^{m}I\left(\hat{q}_j^{(r)}=q_j\right),\qquad(18)$$

其中，r表示500次重复模拟实验中的第r次重复实验，$\hat{q}_j^{(r)}$表示第r次重复模拟中新题j的q向量估计值，q_j表示新题j的q向量真值。$I(\hat{q}_j^{(r)}=q_j)$为指示性函数，用于评估第r次重复模拟中$\hat{q}_j^{(r)}$是否等于q_j。AVCER值越大，新题Q矩阵估计正确率越高。

近似均方根误差(Root Mean Squared Error, RMSE) RMSE指标用于评价新题项目参数的估计正确性，其表达式可写为：

$$RMSE=\sqrt{\frac{1}{500}\sum_{r=1}^{500}\frac{1}{2M}\left(\sum_{j=1}^{M}(\hat{s}_j^r-s_j)^2+\sum_{j=1}^{M}(\hat{g}_j^r-g_j)^2\right)}, \tag{19}$$

上式中，$\hat{s}_j^{(r)}$和$\hat{g}_j^{(r)}$分别表示第r次重复模拟中，新题j的失误参数s和猜测参数g估计值，s_j和g_j分别表示新题j的失误参数s和猜测参数g真值。RMSE值越小，项目参数的估计精度越高。

标定效率：即平均运行时间(Average Running Time, ART) ART用于评估各在线标定方法的标定效率，其计算如下：

$$ART=\frac{\sum_{r=1}^{500}t_r}{500}, \tag{20}$$

其中，t_r表示第r次重复模拟中，各在线标定方法标定新题所用的时间。ART值越小，用于标定新题的方法的效率越高。

4.2 实验结果

图1、表2和图2分别呈现了标定方法SIE、SIE-R-BIC、RMSEA-N和IGEOCM的项目标定精度以及标定效率结果。根据Chen等人(2015)的研究，两方法间标定精度的差值大于等于1%表明一种方法优于另一种方法。总体而言，IGEOCM具有较好的项目标定精度和估计效率，其性能整体上优于SIE、SIE-R-BIC和RMSEA-N方法。由图1可知，IGEOCM的Q矩阵估计正确率高于其它三种方法，属性掌握模式为高阶分布和正态分布时，各方法间的差异更为明显。如在属性掌握模式为均匀分布时，SIE方法和IGEOCM间的最大AVCER差值为2.3%，而在属性掌握模式为高阶分布和正态分布时，两方法间的最大AVCER差值分别高达6.8%和9.1%。

SIE和SIE-R-BIC方法的Q矩阵标定精度在各条件下均较为接近，而RMSEA-N方法在高阶分布和正态分布下的Q矩阵标定正确率低于SIE和SIE-R-BIC方法。在属性掌握模式分布对Q矩阵标定精度的影响上，SIE、SIE-R-BIC、RMSEA-N和IGEOCM的Q矩阵估计正确率在属性掌握模式为均匀分布时最好，高阶分布时次之，正态分布时最差。例如，IGEOCM在均匀、高阶和正态分布下的Q矩阵估计正确率范围分别为80.9%~99.8%，67.0%~97.3%和46.0%~76.7%；而SIE方法在均匀、高阶和正态分布下的Q矩阵估计正确率范围分别为79.0%~99.8%，60.7%~96.9%和38.4%~68.3%。标定样本对各在线标定方法的Q矩阵估计正确率影响较大，标定样本越大，各方法的Q矩阵估计正确率越高。当标定样本$n_j = 40$时，SIE、SIE-R-BIC、RMSEA-N和IGEOCM的平均AVCER值分别为59.6%、60.0%、45.6%和65.1%，而当标定样本$n_j = 200$时，4种方法的平均AVCER值上升到88.1%、88.2%、77.2%和91.2%。因此，增加标定样本可提高各在线标定方法的Q矩阵估计正确率。SIE、SIE-R-BIC、RMSEA-N和IGEOCM方法在被试作答新题个数为4、6和8的情况下均具有相近的Q矩阵估计正确率。

表2为SIE、SIE-R-BIC、RMSEA-N和IGEOCM的项目参数标定结果。SIE方法和IGEOCM在项目参数标定精度上具有相似的性能，其最大RMSE差值不超过0.2%，大多数实验条件下两方法的RMSE值相等。SIE-R-BIC方法的RMSE值在标定样本较少时略低于SIE方法和IGEOCM (如，$n_j = 40$)，在标定样本较多时略高于SIE方法和IGEOCM (如，$n_j = 200$)；RMSEA-N方法的RMSE值在多数条件下都高于SIE、SIE-R-BIC和IGEOCM。在属性掌握模式分布对项目参数标定精度的影响上，SIE、SIE-R-BIC和IGEOCM的项目参数标定精度在属性掌握模式为高阶分布时最好，而RMSEA-N的项目参数标定精度在属性掌握模式为均匀分布时最好。如IGEOCM在高阶、均匀和正态分布下的平均RMSE值分别为0.056、0.066和0.071，RMSEA-N在高阶、均匀和正态分布下的平均RMSE值分别为0.093、0.088和0.142。各方法的项目参数标定精度随标定样本的增加而提升。如，标定样本$n_j = 40$时，SIE方法和IGEOCM的平均RMSE值均为0.11，而当标定样本$n_j = 200$时，两方法的平均RMSE值均减少为0.04。与Q矩阵标定精度一致，被试作答新题个数对SIE、

SIE-R-BIC、RMSEA-N和IGEOCM方法项目参数标定精度的影响可忽略不计。

图 2 为使用 SIE、SIE-R-BIC、RMSEA-N 和 IGEOCM 方法估计 24 个新题的平均运行时间。各模拟条件下，4 种在线标定方法均使用 *R*4.0 运行，其计算机配置相同(如 Intel Core i5-8400 2.81GHz，内存 20G)，因此各标定方法的估计效率具有可比性。由图 2 结果可知，相比于 IGEOCM，SIE、SIE-R-BIC 和 RMSEA-N 方法的估计效率更低，其所有条件下的平均 ART 值约为 IGEOCM 的 49 倍。属性掌握模式分布与被试作答新题个数对 SIE、SIE-R-BIC、RMSEA-N 和 IGEOCM 的估计效率影响较小。此外，SIE、SIE-R-BIC、RMSEA-N 和 IGEOCM 的平均运行时间均随标定样本的增加而延长。当标定样本$n_j = 40$时，SIE、SIE-R-BIC、RMSEA-N 和 IGEOCM 的平均 ART 值分别为 106.22、93.38、61.39 和 1.74，而当标定样本$n_j = 200$时，4 种方法的平均 ART 值延长至 414.71、322.40、286.06 和 6.91。

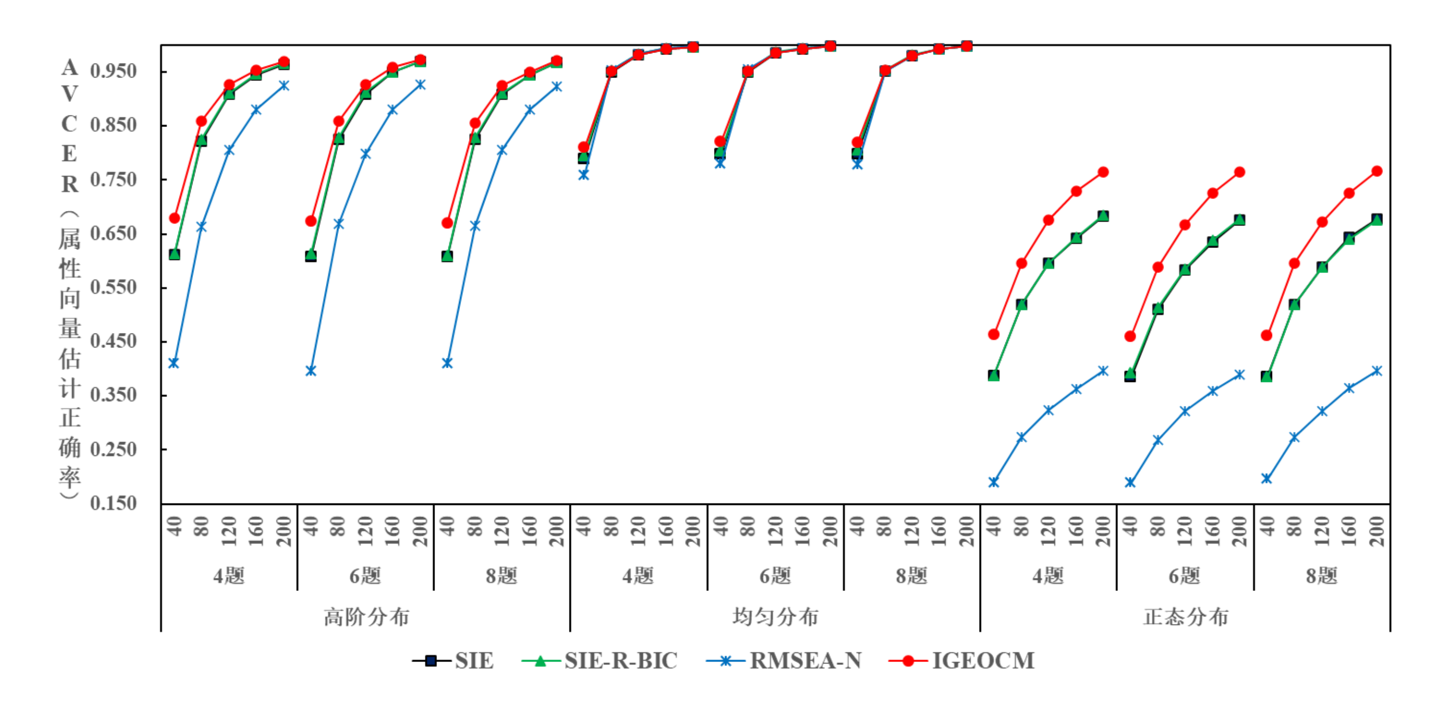


图 1 各在线标定方法在不同条件下的 AVCER(属性向量估计正确率)结果

表 2 各在线标定方法在不同条件下的 RMSE(均方根误差)结果

分布	项目	方法	40	80	120	160	200
高阶	4	SIE	0.090	0.060	0.048	0.041	0.036

		SIE-R-BIC	0.088	0.065	0.057	0.052	0.049
		RMSEA-N	0.132	0.099	0.086	0.079	0.073
		IGEOCM	0.090	0.060	0.048	0.041	0.036
	6	SIE	0.092	0.061	0.049	0.041	0.037
		SIE-R-BIC	0.089	0.066	0.057	0.053	0.050
		RMSEA-N	0.132	0.099	0.085	0.077	0.074
		IGEOCM	0.092	0.061	0.049	0.041	0.037
	8	SIE	0.095	0.060	0.048	0.042	0.037
		SIE-R-BIC	0.090	0.066	0.057	0.053	0.050
		RMSEA-N	0.132	0.098	0.085	0.078	0.074
		IGEOCM	0.095	0.061	0.048	0.042	0.037
均匀	4	SIE	0.123	0.071	0.055	0.046	0.041
		SIE-R-BIC	0.097	0.068	0.057	0.051	0.047
		RMSEA-N	0.118	0.090	0.082	0.078	0.076
		IGEOCM	0.121	0.071	0.055	0.046	0.041
	6	SIE	0.121	0.069	0.053	0.045	0.039
		SIE-R-BIC	0.097	0.068	0.056	0.050	0.046
		RMSEA-N	0.116	0.090	0.081	0.078	0.076
		IGEOCM	0.119	0.069	0.053	0.045	0.039
	8	SIE	0.122	0.071	0.054	0.046	0.040
		SIE-R-BIC	0.097	0.068	0.057	0.051	0.047
		RMSEA-N	0.116	0.090	0.082	0.078	0.076
		IGEOCM	0.121	0.071	0.054	0.046	0.040
正态	4	SIE	0.126	0.076	0.059	0.049	0.044
		SIE-R-BIC	0.099	0.073	0.064	0.058	0.055
		RMSEA-N	0.170	0.149	0.138	0.130	0.123
		IGEOCM	0.126	0.076	0.059	0.049	0.044
	6	SIE	0.124	0.076	0.059	0.050	0.044

	SIE-R-BIC	0.098	0.073	0.064	0.058	0.055
	RMSEA-N	0.171	0.149	0.138	0.129	0.125
	IGEOCM	0.123	0.076	0.059	0.050	0.044
8	SIE	0.129	0.079	0.059	0.049	0.044
	SIE-R-BIC	0.100	0.074	0.063	0.058	0.055
	RMSEA-N	0.170	0.149	0.136	0.128	0.121
	IGEOCM	0.130	0.079	0.060	0.050	0.044

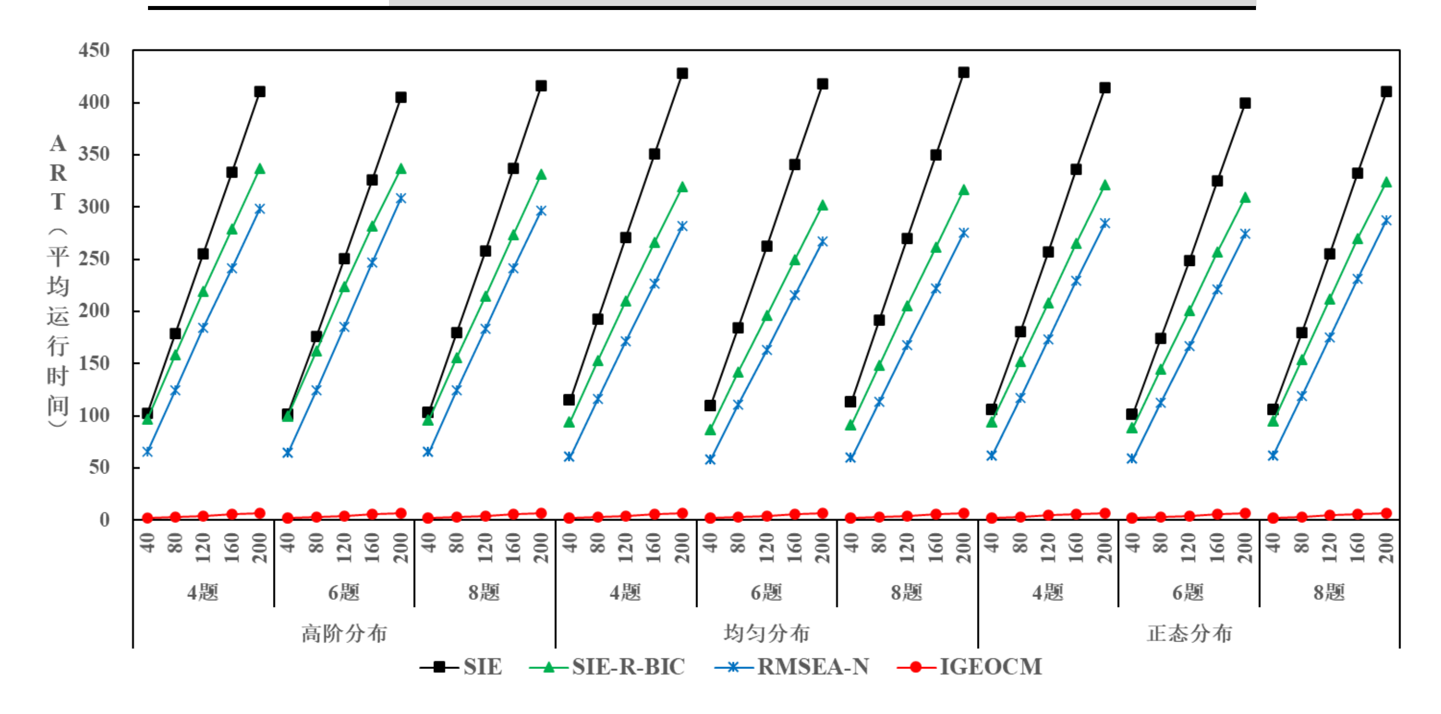


图 2 各在线标定方法在不同条件下的 ART(平均运行时间)结果(单位：秒)

5 研究 2：选题策略对 IGEOCM 和已有在线标定方法性能的影响

IGEOCM、SIE、SIE-R-BIC和RMSEA-N方法基于CD-CAT测验中被试属性掌握模式和属性掌握模式后验分布的估计值以及被试在新题上的作答来标定新题，被试属性掌握模式及属性掌握模式后验分布的估计精度影响各在线标定方法的标定精度(Chen et al., 2015)。而CD-CAT中，选题策略是影响被试属性掌握模式估计精度的重要因素之一。因此，研究2在研究1的基础上，进一步考察选题策略对各在线标定方法性能的影响。

5.1 实验设计

研究2的实验设计和模拟过程与研究1基本一致，但研究2在研究1的基础上新增了MPWKL (the modified PWKL)、GDI (the generalized deterministic inputs, noisy “and” gate (G-DINA) model discrimination index)和香浓熵(Shannon entropy, SHE)选题策略(Cheng, 2009; Kaplan et al., 2015)，以比较IGEOCM和SIE方法在不同选题策略下的可行性和准确性。由于SIE和SIE-R-BIC方法的项目标定精度略高于RMSEA-N方法，且SIE方法的项目参数标定精度在多数条件下均略高于SIE-R-BIC和RMSEA-N方法。另外，三者在标定效率上差异较小，均耗时较长（ART比值不超过1倍），因此研究2中仅选择已有方法SIE作为新方法IGEOCM的比较方法。此外，基于研究1的结果，被试新题作答项目个数对SIE方法和IGEOCM项目标定精度的影响较小，研究2中将被试作答新题的个数固定为6 ($D=6$)。考虑到SIE方法和IGEOCM的运行时间随标定样本的增长而延长，因此研究2中将标定样本固定为40以缩短实验时长。其余实验条件和模拟过程请参见研究1。

5.2 实验结果

表3为SIE方法和IGEOCM在不同选题策略和不同属性掌握模式分布下的项目标定精度与标定效率结果。与研究1结果相似，相比于SIE方法，IGEOCM在各选题策略下均具有更高的项目标定精度和项目估计效率。此外，在所有选题策略下，SIE和IGEOCM方法的Q矩阵估计正确率在属性掌握模式为均匀分布时最好，高阶分布时次之，正态分布时最差。

CD-CAT选题策略对在线标定方法的新题Q矩阵标定精度有一定影响。如属性掌握模式为高阶分布的情况下，SIE方法在选题策略为MPWKL时具有较高的Q矩阵标定精度，其AVCER值为61.7%；SIE方法在选题策略为PWKL时具有较低的Q矩阵标定精度，其AVECR值为60.7%。在属性掌握模式为正态分布的情况下，IGEOCM方法在选题策略为GDI时具有较高的Q矩阵标定精度，其AVCER值为46.7%；IGEOCM方法在选题策略为PWKL时具有较低的Q矩阵标定精度，其AVECR值为45.4%。CD-CAT选题策略对新题项目参数和估计效率的影响可忽略不计。各选题策略下的RMSE均值之差不超过0.2%，平均运行时间(ART)较为接

近。

表 3 SIE 方法和 IGEOCM 在不同条件下的项目标定精度与标定效率结果

分布	方法	AVCER				RMSE				ART			
		PWKL	MPWKL	SHE	GDI	PWKL	MPWKL	SHE	GDI	PWKL	MPWKL	SHE	GDI
高阶	SIE	0.607	0.617	0.615	0.614	0.082	0.083	0.083	0.083	78.438	78.083	78.116	77.818
	IGEOCM	0.678	0.677	0.676	0.679	0.082	0.084	0.082	0.083	1.808	1.811	1.800	1.797
均匀	SIE	0.809	0.807	0.814	0.808	0.089	0.090	0.090	0.089	90.388	89.742	90.421	89.702
	IGEOCM	0.828	0.827	0.831	0.825	0.089	0.090	0.090	0.089	1.861	1.846	1.857	1.845
正态	SIE	0.385	0.383	0.383	0.384	0.099	0.099	0.100	0.099	81.850	81.420	81.752	81.587
	IGEOCM	0.454	0.462	0.457	0.467	0.099	0.099	0.099	0.099	1.884	1.865	1.873	1.880

6 总结与讨论

CD-CAT中同时标定新题Q矩阵和项目参数的在线标定方法较少，且均为参数化的方法，标定新题的时间较长，标定效率较低。因此，研究借鉴数据挖掘中特征选择(Feature Selection)的思路，提出了基于熵的信息增益的在线标定方法(IGEOCM)，以期为CD-CAT题库中项目的增补提供一种更为高效、准确的方法。不同于CD-CAT中已有的在线标定方法，IGEOCM使用非参数的方法标定新题的Q矩阵，较为有效地避免了项目参数估计偏差所带来的影响，改善了项目标定的精度，同时提高了项目标定的效率。随后，使用Monte Carlo模拟研究来验证IGEOCM的可行性和准确性，并将其与已有在线标定方法SIE、SIE-R-BIC和RMSEA-N进行比较。研究结果表明：(1) IGEOCM在各条件下均具有较好的项目标定精度和项目估计效率，且整体上优于SIE、SIE-R-BIC和RMSEA-N方法。SIE等方法基于新题项目参数的估计值来估计新题Q矩阵，项目参数的估计误差影响新题Q矩阵的标定精度，继而降低新题标定精度；而IGEOCM基于被试属性掌握模式及其在新题上的作答直接标定新题Q矩阵，新题Q矩阵的标定与项目参数的估计精度无关，额外影响因素少，新题标定精度更高一些。此外，尽管SIE和IGEOCMSIE项目参数估计方法一致，但SIE方法使用参数化方法标定新题Q矩阵，而IGEOCM方法使用非参数化方法标定新题Q矩阵。相比参数化方法，非参数化

方法计算更为简单，运行时间更短(Chiu et al., 2018)，因此IGEOCM的新题标定效率较好一些。(2) SIE、SIE-R-BIC、RMSEA-N和IGEOCM的项目标定精度随标定样本的增加而提高，4种方法的运行时间随标定样本的增加而延长。(3) SIE、SIE-R-BIC、RMSEA-N和IGEOCM在属性掌握模式分布为均匀分布和高阶分布时的项目标定精度高于正态分布。(4)被试作答新题个数对SIE、SIE-R-BIC、RMSEA-N和IGEOCM项目标定精度和估计效率的影响较小。(5) CD-CAT中选题策略影响SIE方法和IGEOCM的Q矩阵标定精度。在属性掌握模式为高阶和正态分布时，相比PWKL选题策略，SIE方法和IGEOCM在选题策略分别为MPWKL和GDI时的Q矩阵标定精度略高。此外，研究还考察了属性掌握模式为高阶分布时不同的λ_{0k}和λ_{1k}模拟方式对SIE方法和IGEOCM的影响，即λ_{0k}从标准正态中产生和λ_{1k}从对数标准正态分布中产生，在被试作答新题个数固定为6，其余条件与研究1相同的情况下：IGEOCM在该模拟方式下仍优于SIE方法。该结果进一步表明IGEOCM的可行性其及优势（具体数据结果参见附表1）。

当然，研究仍有许多不足之处，今后研究中需加以改进与完善。首先，文中仅验证了所提出IGEOCM在DINA模型下的性能，其在较为复杂的认知诊断模型，如缩减重新参数化融合模型(the Reduced Reparametrized Unified Model, RRUM; Hartz, 2002)，拓广DINA (the Generalized Deterministic Inputs, Noisy and Gate Model, G-DINA; de la Torre, 2011)等模型下的性能仍有待进一步探讨。不同于DINA模型，其仅将被试分为掌握与非掌握两个类别。在更为复杂的模型下，基于被试属性掌握模式和项目q向量可以将被试划分为更多不同的类别，而基于熵的信息增益指标会随着被试所划分类别的增加而增加，因此在更为复杂的认知诊断模型下使用基于熵的信息增益指标来标定新题q向量的效果值得探讨。未来研究中可考虑如何解决被试类别数量对IGEOCM的影响，如对被试类别数进行惩罚以减少类别个数对IGEOCM的影响。

其次，CD-CAT中已有的在线标定方法均是基于二级计分模型。实际上，心理与教育评估中存在大量的多级计分数据以及多级计分题目，且相比于二级计分的作答数据，多级计分的作答数据可为被试提供更为全面详尽的诊断信息。文中所提出的在线标定方法应如何推广到系列G-DINA模型(sequential G-DINA model; Ma & de la Torre, 2016)等多级计分模型之中，并验证其在多级计分模型下的性能

有待进一步研究。

再次，研究为每个被试随机选择新题，用于标定每个新题的被试可能并非最合适的被试。未来研究中可考虑使用自适应的方法来为每个项目选择最合适的被试，比如使用最优设计准则来为每个项目选择最佳被试(He et al., 2020)。然后考察不同的新题选择方式(随机选择和自适应选择)对在线标定方法的影响。

最后，研究假设测验所测量的属性之间相互独立。然而，在实际的诊断测验中，属性之间可能存在各种层级关系，比如无结构型、线型、分支型和收敛型(Leighton et al., 2004)。因此，未来研究一个可考虑的方向是探讨不同属性层级关系对在线标定方法的影响。另外，研究使用模拟实验验证所提出的Q矩阵与项目参数在线标定方法的科学性与合理性，虽然模拟研究的结果能为实践应用提供一定指导，但模拟研究是在理想的情境下进行，会忽略很多真实情境中的影响因素，因此未来研究需进一步评估真实情境中各在线标定方法的性能。总之，CD-CAT中同时标定新题Q矩阵与项目参数的在线标定方法仍需进一步的研究。

参考文献

Ban, J. C., Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A Comparative Study of On-line Pretest Item—Calibration/Scaling Methods in Computerized Adaptive Testing. *Journal of Educational Measurement*, *38*(3), 191–212.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28.

Chen, J. (2017). A residual-based approach to validate Q-matrix specifications. *Applied Psychological Measurement*, *41*(4), 277–293.

Chen, P. (2016). Two new online calibration methods for computerized adaptive testing. *Acta Psychologica Sinica*, *48*(9), 1184–1198.

[陈平. (2016). 两种新的计算机化自适应测验在线标定方法. *心理学报*, *48*(9), 1184–1198.]

Chen, P. (2017). A Comparative Study of Online Item Calibration Methods in Multidimensional Computerized Adaptive Testing. *Journal of Educational and Behavioral Statistics*, *42*(5), 559–590.

Chen, P., & Wang, C. (2015). A New Online Calibration Method for Multidimensional Computerized Adaptive Testing. *Psychometrika*, *81*(3), 674–701.

Chen, P., Wang, C., Xin, T., & Chang, H. H. (2017). Developing new online calibration methods for multidimensional computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 81–117.

Chen, P., & Xin, T. (2011a). Developing on-line calibration methods for cognitive diagnostic computerized adaptive testing. *Acta Psychologica Sinica*, *43*(6), 710–724.

[陈平，辛涛. (2011a). 认知诊断计算机化自适应测验中在线标定方法的开发. *心理学报*, *43*(6), 710–724.]

Chen, P., & Xin, T. (2011b). Item replenishing in cognitive diagnostic computerized adaptive testing. *Acta Psychologica Sinica*, *43*(7), 836–850.

[陈平，辛涛. (2011b). 认知诊断计算机化自适应测验中的项目增补. *心理学报*, *43*(7), 836–850.]

Chen, P., Xin, T., Wang, C., & Chang, H. (2012). Online Calibration Methods for the DINA Model with Independent Attributes in CD-CAT. *Psychometrika*, *77*(2), 201–222.

Chen, Y., Liu, J., & Ying, Z. (2015). Online item calibration for Q-matrix in CD-CAT. *Applied psychological measurement*, *39*(1), 5–15.

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*(4), 619–632.

Chiu, C. Y., Sun, Y., & Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika*, *83*(2), 355–375.

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199.

de la Torre, J., & Chiu, C. Y. (2016). General method of empirical Q-matrix validation. *Psychometrika*, *81*(2), 253–273.

Fleuret, F. (2004). Fast binary feature selection with conditional mutual information. *Journal of Machine learning research*, *5*(11), 1531–1555.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*(3), 1157–1182.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality* (Unpublished doctoral dissertation). University of Illinois at Urbana–Champaign.

He, Y., Chen, P., & Li, Y. (2020). New Efficient and Practicable Adaptive Designs for Calibrating Items Online. *Applied Psychological Measurement*, *44* (1), 3–16.

Hoque, N., Bhattacharyya, D. K., & Kalita, J. K. (2014). MIFS-ND: A mutual information-based feature selection method. *Expert Systems with Applications*, *41*(14), 6371-6385.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272.

Kaplan, M., de la Torre, J., & Barrada, J. R. (2015). New Item Selection Methods for Cognitive Diagnosis Computerized Adaptive Testing. *Applied Psychological Measurement*, *39*(3), 167–188.

Lee, S., Park, Y. T., & d'Auriol, B. J. (2012). A novel feature selection method based on normalized mutual information. *Applied Intelligence*, *37*(1), 100–120.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The Attribute Hierarchy Method for Cognitive Assessment: A Variation on Tatsuoka's Rule-Space Approach. *Journal of Educational Measurement*, *41*(3), 205–237.

Li, H. (2012). *Statistical learning method*. Beijing: Tsinghua University Press.

[李航. (2012). *统计学习方法*. 北京：清华大学出版社.]

Liu, H., You, X., Wang, W., Ding, S., & Chang, H. (2013). The Development of Computerized Adaptive Testing with Cognitive Diagnosis for an English Achievement Test in China. *Journal of Classification*, *30*(2), 152–172.

Ma, W., & de la Torre, J. (2016). A sequential cognitive diagnosis model for polytomous responses. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 253–275.

Pereira, R. B., Plastino, A., Zadrozny, B., & Merschmann, L. H. (2015). Information gain feature selection for multi-label classification. *Journal of Information and Data Management*, *6*(1), 48–58.

Rupp, A. A., & Templin, J. L. (2008). The effects of Q-matrix misspecification on parameter estimates and classification accuracy in the DINA model. *Educational and Psychological Measurement*, *68*(1), 78–96.

Stocking, M. L. (1988). Scale Drift in On-Line Calibration. *Ets Research Report*, *1988*(1), 1–122.

Tan, Q. (2019). *The Development of Generalized Online Calibration Methods in CD-CAT* (Unpublished master's thesis). Jiangxi Normal University, Nanchang.

[谭青蓉. (2019). *CD-CAT 广义在线标定方法开发研究* (硕士学位论文). 江西师范大学，南昌.]

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer (Ed), *Computerized adaptive testing*: *A primer* (Chap. 4, pp. 65–102). Hillsdale, NJ: Erlbaum.

Wang, C. (2013). Mutual Information Item Selection Method in Cognitive Diagnostic Computerized Adaptive Testing with Short Test Length. *Educational and Psychological Measurement*, *73*(6), 1017–1035.

Weiss, D. J. (1982). Improving Measurement Quality and Efficiency with Adaptive Testing. *Applied Psychological Measurement*, *6*(4), 473–492.

Yu, X., & Cheng, Y. (2020). Data-driven Q-matrix validation using a residual-based statistic in cognitive diagnostic assessment. *British Journal of Mathematical and Statistical Psychology*, *73* (Suppl), 145–179.

A High-efficiency and New Online Calibration Method in CD-CAT Based on Information Gain of Entropy and EM algorithm

TAN Qingrong[1], WANG Daxun[1], LUO Fen[1], CAI Yan[1], TU Dongbo[1]

([1]*School of Psychology, Jiangxi Normal University, Nanchang 330022, China*)

Abstract

Cognitive diagnostic computerized testing (CD-CAT) includes the advantages of both cognitive diagnosis (CD) and computerized adaptive testing (CAT), which can offer detailed diagnosis feedback for each examinee by applying fewer test items and time. It has been a promising field. An item bank is a prerequisite for the implementation of CD-CAT. However, its maintenance is a very challenging task. One of the effective ways to maintain the item bank is online calibration. Till now, there are only a few online calibration methods in the CD-CAT context that can calibrate Q-matrix and item parameters simultaneously. Moreover, the computational efficiency of these methods needs to be further improved. Therefore, it is crucial to find more online calibration methods that jointly calibrate the Q-matrix and item parameters.

Inspired by the SIE (Single-Item Estimation) method proposed by Chen et al. (2015) and information gain feature selection criteria in feature selection, an information gain of entropy-based online calibration method (IGEOCM) was proposed in this study. The proposed method can jointly calibrate Q-matrix and item parameters in a sequential manner. The detailed process of the new items was described as follows: First, for the new item *j*, the q-vector can be calibrated by maximizing the information gain of entropy-based on the basis of the attribute patterns of examinees and the examinees' responses to item *j*. Second, the item parameters of the new item *j* are estimated by the EM algorithm based on the posterior distribution of examinees' attribute pattern, the examinees' responses to item *j*, and the q-vector estimated in the first step. The first step and second step are repeated for all other new items to obtain their estimated Q-matrix and estimated item parameters item by item. Two simulation studies were conducted to examine whether the IGEOCM could accurately and efficiently calibrate the Q-matrix and item parameters of the new items under different calibration sample sizes (40, 80, 120, 160, and 200), different attribute pattern distributions (uniform distribution, higher-order distribution, and multivariate normal distribution), the different

number of new items answered by examinee (4, 6, and 8), and different item selection algorithms (posterior-weighted Kullback-Leibler, PWKL; the modified PWKL, MPWKL; the generalized deterministic inputs, noisy "and" gate model discrimination index, GDI; and Shannon entropy, SHE). Furthermore, the performance of the proposed methods was compared with the SIE, SIE-R-BIC, and RMSEA-N methods.

The results indicated that (1) The IGEOCM worked well in terms of the calibration accuracy and estimation efficiency under all conditions, and outperformed the SIE, SIE-R-BIC, and RMSEA-N methods overall. (2) The accuracy of the item calibration increases as the sample size increases for all calibration methods under all conditions. (3) The SIE, SIE-R-BIC, RMSEA-N, and IGEOCM performed better under the uniform distribution and higher-order distribution than under the multivariate normal distribution. (4) The number of new items answered by the examinee had a negligible impact on the calibration accuracy and computation efficiency of the SIE, SIE-R-BIC, RMSEA-N, and IGEOCM. (5) The item selection algorithm in CD-CAT affects the Q-matrix calibration accuracy of the SIE and IGEOCM methods. Under the higher-order distribution and multivariate normal distribution, the SIE method and IGEOCM had higher Q-matrix calibration accuracy when the item selection algorithms were MPWKL and GDI.

On the whole, although the proposed IGEOCM is competitive and outperforms the conventional method irrespective of the calibration precision or computational efficiency, the studies on the online calibration method in CD-CAT still need to be further deepened and expanded.

Keywords Cognitive diagnostic computerized adaptive testing, Item replenishing, Online calibration, Q-matrix, Information gain of entropy

附录：

【例1】：假设被试的属性掌握模式已知且呈均匀分布，被试在新题j上的作答不存在失误和猜测。值得注意的是，由于作答新题j的被试人数n_j以及每个被试在新题j上的作答在所有可能的q向量下都是固定不变的，熵的信息增益$g(R_j,q_j)$中作答数据集R_j的熵值$E(R_j)$在所有可能q向量下都相等，其大小完全取决于条件熵$E(R_j|q_j)$的大小。因此，本例中主要通过描述不同q向量下条件熵$E(R_j|q_j)$的改变来说明不同q向量下熵的信息增益$g(R_j,q_j)$的改变。令测验测量的属性个数K=3，则共有$2^K=8$种可能的属性掌握模式，每一种属性掌握模式的期望人数为$n_j/8$。若新题j的正确q向量为$q_j^{correct}=[1\ 0\ 0]$，对于DINA模型，则属性掌握模式为[1 0 0]、[1 1 0]、[1 0 1]和[1 1 1]的被试将被划分为掌握组（h_1），而属性掌握模式为[0 0 0]、[0 1 0]、[0 0 1]和[0 1 1]的被试将被划分为非掌握组（h_0）。由于被试的属性掌握模式为均匀分布，掌握组h_1和非掌握组h_0中的被试人数均为$n_j/2$（$n_{jh_1}=n_{jh_0}=4\times n_j/8=n_j/2$）。若$h_1$中被试在项目$j$上作答错误和作答正确的人数分别为$n_{j10}$和$n_{j11}$，$h_0$中被试在项目$j$上作答错误和作答正确的人数分别为$n_{j00}$和$n_{j01}$，则$n_{jh_1}=n_{j10}+n_{j11}$和$n_{jh_0}=n_{j00}+n_{j01}$。将$h_1$中每种属性掌握模式的被试在项目$j$上作答错误的人数分别标记为$F_1,F_2,\cdots,F_4$，$h_0$中每种属性掌握模式的被试在项目$j$上作答正确的人数分别标记为$T_1,T_2,\cdots,T_4$。则

$$E\left[\frac{n_{j10}}{n_{jh_1}}\right]=2\times\frac{E[F_1]+E[F_2]+\cdots+E[F_4]}{n_j},$$

$$E\left[\frac{n_{j01}}{n_{jh_0}}\right]=2\times\frac{E[T_1]+E[T_2]+\cdots+E[T_4]}{n_j}.$$

基于被试在新题j上的作答不存在失误和猜测这一假设，掌握组h_1中每种属性掌握模式下的被试在新题j上答错的期望人数为0，即$E[F_1]=E[F_2]=\cdots=E[F_4]=0$；非掌握组$h_0$中每种属性掌握模式下的被试在新

题 j 上答对的期望人数也为 0，即 $E[T_1]=E[T_2]=\cdots=E[T_4]=0$。且在被试在新题 j 上的作答不存在失误和猜测的情况下，$\frac{n_{j10}}{n_{jh_1}}=E\left[\frac{n_{j10}}{n_{jh_1}}\right]$ 和 $\frac{n_{j01}}{n_{jh_0}}=E\left[\frac{n_{j01}}{n_{jh_0}}\right]$（Chiu et al., 2018）。则有，

$$\frac{n_{j10}}{n_{jh_1}}=\frac{2\times4\times0}{n_j}=0,\qquad \frac{n_{j11}}{n_{jh_1}}=1-\frac{n_{j10}}{n_{jh_1}}=1.$$

$$\frac{n_{j01}}{n_{jh_0}}=\frac{2\times4\times0}{n_j}=0,\qquad \frac{n_{j00}}{n_{jh_0}}=1-\frac{n_{j01}}{n_{jh_0}}=1.$$

然后 $E(R_j|q_j^{correct})$ 可计算如下

$$\begin{aligned}E(R_j|q_j^{correct})&=\left[\frac{n_{jh_1}}{n_j}\times\left[\left(\frac{n_{j10}}{n_{jh_1}}\times\log\frac{n_{j10}}{n_{jh_1}}\right)+\left(\frac{n_{j11}}{n_{jh_1}}\times\log\frac{n_{j11}}{n_{jh_1}}\right)\right]\right]+\frac{n_{jh_0}}{n_j}\times\left[\left(\frac{n_{j00}}{n_{jh_0}}\times\log\frac{n_{j00}}{n_{jh_0}}\right)+\left(\frac{n_{j01}}{n_{jh_0}}\times\log\frac{n_{j01}}{n_{jh_0}}\right)\right]\\&=\left[\frac{1}{2}\times[(0\times\log0)+(1\times\log1)]\right]+\left[\frac{1}{2}\times[(0\times\log0)+(1\times\log1)]\right]\\&=0\end{aligned}$$

.

若新题 j 的 q 向量错误 $q_j^{incorrect}=[0\ \ 1\ \ 1]$，则属性掌握模式为[0 1 1]和[1 1 1]的被试将被划分为 h_1，而属性掌握模式为[0 0 0]、[1 0 0]、 [0 1 0]、[0 0 1]、[1 1 0]和[1 0 1]的被试则将被划分为 h_0。h_1 和 h_0 中的被试人数分别为 $n_j/4$（$n_{jh_1}=2\times n_j/8=n_j/4$）和 $3n_j/4$（$n_{jh_0}=6\times n_j/8=3n_j/4$）。此时，错误的 q 向量将原本在新题 j 上作答正确的掌握组被试错误地分入非掌握组，被错分入非掌握组的被试的属性掌握模式为[1 0 0]、[1 1 0]和[1 0 1]，这些模式的被试在新题 j 上作答正确的期望人数为 $E[T_2]=E[T_5]=E[T_6]=n_j/8$；而将原本在新题 j 上作答错误的非掌握组被试错误地分入掌握组，被错分入掌握组的被试的属性掌握模式为[0 1 1]，该模式中被试在新题 j 作答错误的期望人数为 $E[F_1]=n_j/8$。因此，

$$\frac{n_{j10}}{n_{jh_1}}=\frac{4\times\left[(1\times0)+\left(1\times\frac{n_j}{8}\right)\right]}{n_j}=0.5.\qquad \frac{n_{j11}}{n_{jh_1}}=1-\frac{n_{j10}}{n_{jh_1}}=0.5.$$

$$\frac{n_{j01}}{n_{jh_0}}=\frac{4\times\left[(3\times0)+\left(3\times\frac{n_j}{8}\right)\right]}{3n_j}=0.5.\qquad \frac{n_{j00}}{n_{jh_0}}=1-\frac{n_{j01}}{n_{jh_2}}=0.5.$$

然后$E(R_j|q_j^{incorrect})$可计算如下

$$E(R_j|q_j^{incorrect}) = \left[-\frac{1}{4}\times[(0.5\times\log 0.5)+(0.5\times\log 0.5)]\right]+\left[-\frac{3}{4}\times[(0.5\times\log 0.5)+(0.5\times\log 0.5)]\right] = 0.69$$

.

由上述例子可知，在新题j的q向量正确时，$E(R_j|q_j^{correct})$最小，其值为 0，此时熵的信息增益$g(R_j, q_j)$达到最大。因此，在新题q向量未知的情况下，可以选择能使熵的信息增益$g(R_j, q_j)$最大的q向量作为新题j的估计q向量。

附表 1 不同λ_{0k}和λ_{1k}产生方式下 SIE 和 IGEOCM 方法的项目标定精度

		方法	40	80	120	160	200
AVCER	条件 1	SIE	0.589	0.786	0.860	0.897	0.920
		IGEOCM	0.641	0.823	0.885	0.913	0.938
	条件 2	SIE	0.606	0.812	0.896	0.942	0.965
		IGEOCM	0.668	0.857	0.916	0.950	0.966
RMSE	条件 1	SIE	0.134	0.088	0.068	0.058	0.051
		IGEOCM	0.132	0.085	0.069	0.060	0.052
	条件 2	SIE	0.095	0.062	0.049	0.041	0.037
		IGEOCM	0.095	0.062	0.049	0.041	0.037

注：条件 1 表示λ_{0k}和λ_{1k}分别从正态分布和对数正态分布中产生；条件 2 表示设置$\lambda_0 = (-1, -0.6, -0.2, 0.2. 0.6, 1)$，且对于所有属性$k$均有$\lambda_{1k} = 1.5$。